An Empirical Benchmark for Decadal Forecasts of Global Surface Temperature Anomalies

Matthew Newman

CIRES Climate Diagnostics Center, University of Colorado, and Physical Sciences Division/NOAA Earth System Research Laboratory, Boulder, Colorado

Email: matt.newman@noaa.gov

Journal of Climate, expedited, submitted 7/31/2012

ABSTRACT

The suitability of an empirical multivariate AR1 model as a benchmark for the skill of decadal surface temperature forecasts is demonstrated. Constructed from the observed simultaneous and one-year lag correlation statistics of 12-month running mean sea surface temperature (SST) and surface (2m) land temperature global anomalies for the years 1900-2008, the empirical model hindcasts have skill for leads 2-5 and 6-9 years comparable to and sometimes even better than the CMIP5 model hindcasts initialized annually over the period 1960-2000, and are much more skillful than damped persistence (e.g., a local univariate AR1 process). The pronounced similarity in geographical variations of skill between the empirical model and CMIP5 hindcasts suggests similarity in their sources of skill as well, supporting additional evaluation of the empirical model's skill and predictability over the entire record. It is shown that for forecast leads greater than about a year, the empirical model skill is almost entirely due to patterns corresponding to the secular trend and to two global patterns that each have about ten year decorrelation time scales. In the Atlantic, all three patterns contribute to forecast skill of the Atlantic Multidecadal Oscillation (AMO) index. In the Pacific, only one pattern contributes to the relatively modest long-lead forecast skill of the Pacific Decadal Oscillation (PDO) index, consistent with earlier findings that found an independent decadal signal in the PDO as a residual after both interannual and decadal ENSO influences were first removed. These results suggest that multivariate red noise rather than univariate red noise is the most appropriate baseline comparison for coupled model decadal forecasts.

1. Introduction

The decadal prediction problem has been in an embryonic stage for decades. To progress, we could simply apply the climate community's long experience in understanding seasonal-to-interannual variability and improving its prediction to the decadal variability and prediction problem. For example, a wide array of both physical and empirical methods has been used to make ENSO forecasts (e.g., review by Latif et al. 1998). Statistical forecasts can complement those from physical models, as they are relatively easy and economical to perform, and can be as skillful as physical models for some applications, including ENSO-related forecasts.

It seems reasonable then that a similar two-pronged approach of physical and empirical methods could advance decadal prediction. This is not to say that this improvement will or can occur as readily as was done for seasonal forecasts. One concern is that, while on interannual time scales ENSO provides a very well defined phenomeon that may be understood as the result of a defined mechanism (e.g. delayed oscillator theory, recharge-discharge mechanism), there does not appear to be so clearly a defined decadal "phenomenon", at least in the Pacific. Large scale patterns such as the Pacific decadal oscillation (PDO; Mantua et al. 1997) do not dominate decadal variability to the same degree as ENSO dominates interannual variability, and moreover may represent the superposition and/or convolution of a few mechanisms (e.g., Schneider and Cornuelle 2005; Newman 2007) rather than the result of one identifiable physical process. If most decadal variability represents the low-frequency or reddened tail of interannual phenomena (e.g., Newman et al 2003b; Vimont 2005) rather than truly "decadal" phenomena, then decadal forecasts will likely have very limited predictability. The

effects of anthropogenic climate change complicate comparison between models and observations, and how to distinguish natural decadal variability from anthropogenically-forced decadal variation is a fundamental problem (Solomon et al. 2011).

Currently a number of modeling centers have carried out a series of decadal "hindcasts" as part of the CMIP5 effort (Taylor et al. 2012). It is an important long-range goal of climate diagnosis to provide insights that will help improve decadal forecasts from these CGCMs. Here, we explore the utility of diagnosing annual to decadal variability and predictability in an empirically determined model of the observed system.

2. Multivariate red noise

Climate variability is often characterized by a notable separation between the dominant time scales of interacting processes. For example, compared to much longer ocean timescales, weather varies so rapidly that it has almost no memory. Weather forcing of the ocean can then be approximated as white noise forcing of a damped integrator. This is an example of univariate red noise for an anomaly scalar time series, the simplest null hypothesis for both atmospheric and oceanic climate. When extended to the more general case of anomalies representing many evolving regional patterns of climate variables, this approximation based on time scale separation becomes *multivariate red noise*. As opposed to its univariate counterpart, multivariate red noise represents evolution of both stationary and propagating anomaly patterns (so that scalar indices derived from it can have spectral peaks) and allows for non-symmetric dynamical relationships (so that despite the lack of exponential modal instability, some anomalies experience significant but transient growth and evolution over finite time intervals).

The empirical technique determining multivariate red noise from observations, called linear inverse modeling (LIM), provides an excellent approximation of observed Pacific SST anomaly evolution on time scales ranging from weeks to years. In our prior study (Newman 2007; hereafter N07), we constructed such an empirical model to diagnose forecast skill and predictability of tropical and North Pacific SSTs and found that the empirical model reproduced observed tropical-North Pacific relationships on decadal time scales better than most CMIP3 coupled GCMs. Subsequent studies have had similar success in the Atlantic (Hawkins and Sutton 2009, Zanna 2012) and in both ocean basins (Vimont 2012).

In this paper, the N07 analysis is extended to a state vector constructed from both Pacific and Atlantic SSTs and global surface land temperatures. The empirical model is shown to have skill comparable to three CMIP5 decadal hindcast models that used yearly start dates for the period 1960-2000. The sources of this skill are diagnosed and evaluated in the context of simpler climate indices.

3. Data and model details

Datasets used in this study were SSTs from the Hadley Sea Ice and Sea Surface Temperature analysis (HadISST; Rayner et al. 2003) and surface land temperatures from the University of East Anglia Climatic Research Unit (CRU) TS 3.1 dataset (Mitchell and Jones 2005), both over the period 1900–2009. Monthly data were interpolated onto 2º latitude x 5º longitude gridboxes. Data were temporally smoothed with a 12-month running mean; anomalies were then determined by removing the climatological monthly mean. This allows an analysis that does not consider seasonality. However, seasonality is likely still relevant to decadal variability (e.g., Vimont 2005). Data was prefiltered in an

EOF space that retained about 78% of the SST variance in both the IndoPacific and the Atlantic basins, and about 62% of the surface land temperature variance.

A multivariate AR1 process for a state vector $\mathbf{x}$ can be expressed as

$$\mathbf{x}(t+1) = \mathbf{G}_1\mathbf{x}(t) + \sigma(t), \quad (1)$$

which is the integrated solution of the dynamical system

$$\frac{d\mathbf{x}}{dt} = \mathbf{L}\mathbf{x} + \boldsymbol{\xi} \quad (2)$$

forced by white noise $\boldsymbol{\xi}$, where $\mathbf{G}_1 = exp\,(\mathbf{L})$. N07 determined (2) for Pacific SSTs and we have likewise determined it for global SSTs (not shown). When including surface land temperatures, however, some time scales in $\mathbf{L}$ are too short to be sampled at 1-year intervals, so in this paper we take the simpler route of using (1), solving for $\mathbf{G}_1$ via multiple linear regression. Note, however, that (2) also implies that the best forecast $\hat{\mathbf{x}}(n)$ from initial conditions $\mathbf{x}(0)$ for a lead of n years is

$$\hat{\mathbf{x}}(n) = [\mathbf{G}_1]^n\mathbf{x}(0), \quad (3)$$

and that the lag covariance statistics of $\mathbf{x}$ for a lag of n years is

$$\mathbf{C}(n) = [\mathbf{G}_1]^n\mathbf{C}(0) \quad (4)$$

where $\mathbf{C}(n) = \left\langle \mathbf{x}(t+n)\mathbf{x}(t)^T \right\rangle$ and $\mathbf{C}(0) = \left\langle \mathbf{x}(t)\mathbf{x}(t)^T \right\rangle$. This allows us to still make forecasts using the empirical model and to test its overall validity.

The leading 8/6 EOFs of anomalous IndoPacific/Atlantic SSTs between 60ºS and 60ºN and the leading 6 EOFs of anomalous surface land temperatures were retained for the model. The time-varying coefficients of these EOFs, i.e., the principal components (PCs), define a 20-component state vector $\mathbf{x}$.

Finally, the LIM must be tested on data independent of that used to determine $\mathbf{G}_1$. Estimates of $\mathbf{G}_1$ and of forecast skill were cross-validated as follows. We sub-sampled the data record by removing 10% of the data, calculate $\mathbf{G}_1$ from the remaining 90%, and then generated forecasts for the independent period. This procedure was repeated for all months. All measures of forecast skill in this study are based upon these jack-knifed forecasts; note also that forecasts are compared with the complete (that is, untruncated in EOF space) gridded observations.

Hindcasts from the empirical model are compared to hindcasts from three CMIP5 CGCMs: HadCM3 (DePreSys), MPI-ESM-LR, and GFDL-CM2p1. These models were chosen since they were the only available models whose hindcasts were initialized yearly rather than every five years. Skill was determined from the ensemble mean for each hindcast initialization.

4. Results

Testing the empirical model

We first test the ability of the empirical model to reproduce the lag-covariability statistics of **x**. Figure 1 shows the observed lag-autocovariance for n = 2, 4, 6, and 8 years compared to that predicted by (4). Generally, the match is quite good, and confirms that the empirical multivariate AR1 model represents the statistics of evolving surface temperature anomalies over the 20th century quite well. Note that the empirical model has covariability over land that tends to become too strong, however. Also, while the empirical model captures the anti-correlation in the tropical eastern Pacific due to ENSO for a lag of 2 years, it is a little weak. This is likely a consequence of using only SST to

150 determine the oceanic portion of the state vector. A LIM that explicitly includes
151 subsurface physical processes in its state vector will better reproduce the time evolution
152 of SST anomalies and their statistics, especially for time scales of a year or more
153 (Newman et al. 2011). Still, the empirical model does *implicitly* include those subsurface
154 effects that are linearly related to SST. This is an important distinction from a physical
155 dynamical model in which the evolution of the state vector is governed only by the
156 *explicitly* represented interactions among its components.

157 *Forecast skill of the empirical model and CMIP5 decadal hindcasts*

158 Figure 2 shows forecast skill as measured by local anomaly correlation for forecasts
159 averaged over leads of 2-5 (left panels) and 6-9 (right panels) years. The top two rows,
160 comparing skill from the multivariate AR1 model to that obtained from damped
161 persistence determined locally (i.e., a univariate AR1 model), show that the multivariate
162 AR1 model sets a much higher benchmark for skill. Additionally, the multivariate AR1
163 decadal hindcast skill for yearly start dates from 1960-2005 is comparable to and
164 sometimes better than skill from decadal hindcasts in the CMIP5 archive. Notably, areas
165 of relatively high and relatively low skill often coincide between both the empirical and
166 CGCM hindcasts. LIM can thus serve as a benchmark for decadal hindcast skill.

167 The dependence of skill on forecast lead time for the Atlantic Multidecadal
168 Oscillation (AMO) and PDO indices is shown in Fig. 3. Here, the AMO index is defined
169 as the area-weighted average of North Atlantic SST (between 0ºN and 70ºN) and the
170 PDO index is defined as the projection of SST on the leading EOF of monthly detrended
171 North Pacific SST anomalies (between 20ºN and 70ºN). AMO skill is generally higher
172 than PDO skill, which drops off very rapidly for leads greater than a year. N07 had a

similar result and suggested that it was due to the dependence of the PDO on ENSO, which itself is predictable for only about a year. The multivariate AR1 model again provides a more stringent decadal forecast test than does persistence: for the AMO, differences between GCM and empirical model skill are small and not significant, and for the PDO the empirical model has higher (albeit modest) skill than all three GCMs for leads greater than about 5 years.

Actual and expected forecast skill

The geographical variations of forecast skill are generally similar between the empirical model and the CGCMs. This suggests that, despite the very great differences in model reconstruction, the sources of forecast skill for CGCMs are largely the same as for the multivariate AR1 model.

One of the attractive aspects of the multivariate AR1 approach is that its low order and simplicity makes it a straightforward tool for assessing and diagnosing overall decadal predictability of surface temperatures. It can be shown that for an infinite ensemble forecast skill measured by the average anomaly correlation $\rho_\infty(n)$ between forecast and verification anomalies is also a function of S, the forecast signal-to-noise ratio at lead time n:

$$\rho_\infty(n) = S \Big/ [1 + S^2]^{1/2} \tag{5}$$

(Sardeshmukh et al. 2000). In (1) and (2) we assume noise is independent of the state; so on average S is directly related to stronger predictable signal determined from (3) (Newman et al. 2003a). Therefore, long-range forecasts have highest skill for those states with relatively large initial amplitude in the least-damped eigenmodes of $\mathbf{G}_1$.

195 The validity of (5) is investigated in Fig. 4 where the actual hindcast skill from the
196 entire record (top panels) is compared to the expected skill ρ_∞ (middle panels) for years
197 2-5 and 6-9. In both cases the actual skill, while generally somewhat less than ρ_∞, has a
198 pattern that is very similar to the expected skill. Certainly, while the multivariate AR1
199 model may be a good model of variability of **x**, it is not a perfect one. Also, practical
200 limitations to the empirical determination of $\mathbf{G}_1$ (such as data quality concerns) could be
201 expected to produce errors both in model formulation and model forecasts. For both these
202 reasons, treating the LIM as if it were a perfect model underestimates the actual forecast
203 error. Still, the overall picture suggests that the actual skill is related to variations in
204 forecast signal strength, as expected.

205 Almost all of the skill, both actual and expected, is based on the two leading
206 eigenmodes of $\mathbf{G}_1$, shown in Fig. 5 along with the associated projection coefficient time
207 series. The leading eigenmode is stationary with a very long e-folding time and clearly
208 represents the global secular trend pattern. The second eigenmode is nominally a
209 propagating mode, but in reality it can be considered as two distinct quasi-stationary
210 patterns since the period is very much greater than the 10-year e-folding time. This
211 eigenmode represents decadal variability, primarily over the Atlantic (most energetic
212 phase) and over the Pacific (least energetic phase). The latter is very similar to the second
213 leading eigenmode of the Pacific-only LIM of N07 (dubbed the "Pacific Multidecadal
214 Fluctuation" or PMF), which represented the residual of the PDO when all ENSO
215 influences were first removed. When the projections of all hindcasts and data on these
216 two eigenmodes are removed, the resulting skill map (bottom panels of Fig. 4) shows that
217 essentially no skill remains.

Finally, Fig. 6 shows the impact of different initial conditions on the skill of the PDO in the empirical model. Several new hindcast datasets were created; for every set a portion of the initial condition of each hindcast was first removed. Note that removing the leading eigenmode (i.e, the trend) has almost no impact on PDO skill. The greatest impact occurs when the PMF phase of the second eigenmode is removed; that is, for forecast leads greater than a year (when ENSO impacts are still important) PDO skill is primarily due to PMF persistence.

5. Concluding Remarks

A multivariate AR1 model, empirically constructed from annually averaged surface temperatures using a one-year lag, has been shown to be a more suitable benchmark for decadal forecasts than is damped persistence. In fact, the empirical model has skill that is comparable to the CGCMs, both in amplitude and in geographical variation, suggesting that the much simpler empirical model can also be used to diagnose sources of forecast skill for both forecast systems.

Virtually all long-range skill from the empirical multivariate AR1 model comes from the two eigenmodes with the longest e-folding times. The leading eigenmode represents the global secular trend pattern while the second eigenmode represents decadal variability. Note that the second eigenmode does not *propagate* with a multidecadal period, but instead has a sufficiently long e-folding time that it varies on a multidecadal timescale. The most notable deficiency in CGCM hindcast skill appears to be related to this eigenmode over the Pacific. It is interesting that the similar PMF eigenmode found in the Pacific-only LIM was poorly simulated in all the CMIP3 pre-industrial control and historical model simulations (N07; Solomon et al. 2011). Whether the global version of

241 the eigenmode continues to be poorly represented by the CMIP5 models, and if so, why,
242 is a subject for further investigation.

243 6. Acknowledgements

244 The author thanks Mike Alexander and Amy Solomon for helpful comments. This work
245 was supported by a grant from NOAA CVP.

246

References

Hawkins, E., and R. Sutton, 2009: Decadal predictability of the Atlantic ocean in a coupled GCM: Forecast skill and optimal perturbations using linear inverse modeling. *J. Climate*, **22(**14), 3960–3978.

Latif, M., and and Coauthors, 1998: A review of the predictability and prediction of ENSO. *J. Geophys. Res*., **103**, 14375-14393.

Mantua, N. J., S. R. Hare, Y. Zhang, J. M. Wallace, and R. Francis, 1997: A Pacific interdecadal climate oscillation with impacts on salmon production. *Bull. Amer. Meteor. Soc.*, **78**, 1069–1079.

Mitchell, T. D., and P. D. Jones, 2005: An improved method of constructing a database of monthly climate observations and associated high-resolution grids. *Int. J. Climatol*. **25**, 2005.

Newman, M., 2007: Interannual to decadal predictability of tropical and North Pacific sea surface temperatures. *J. Climate*, **20**, 2333-2356.

Newman, M., P. D. Sardeshmukh, C. R. Winkler, and J. S. Whitaker, 2003a: A study of subseasonal predictability. *Mon. Wea. Rev*., **131**, 1715-1732.

Newman, M., G. P. Compo, and M. A. Alexander, 2003b: ENSO-forced variability of the Pacific Decadal Oscillation. *J. Climate*, **16**, 3853-3857.

Newman, M., M. A. Alexander, and J. D. Scott, 2011: An empirical model of tropical ocean dynamics. *Climate Dynamics*, **37**, 1823-1841.

Penland, C., and P. D. Sardeshmukh, 1995: The optimal growth of tropical sea surface temperature anomalies. *J. Climate*, **8**, 1999—2024.

269 Rayner, N. A., D. E. Parker, E. B. Horton, C. K. Folland, L. V. Alexander, D. P. Rowell,
270 E. C. Kent, and A. Kaplan, 2003: Global analyses of sea surface temperature, sea ice, and
271 night marine air temperature since the late nineteenth century. *J. Geophy Res*., **108**, 4407,
272 doi:10.1029/2002JD002670.
273 Sardeshmukh, P.D., G. P. Compo, and C. Penland, 2000: Changes in probability
274 associated with El Niño. *J. Climate*, **13**, 4268-4286.
275 Schneider, N., and B. D. Cornuelle, 2005: The forcing of the Pacific Decadal Oscillation.
276 *J. Climate*, **18**, 4355-4373.
277 Solomon, A., and the US CLIVAR Working Group on Decadal Predictability, 2011:
278 Distinguishing the roles of natural and anthropogenically forced decadal climate
279 variability: Implications for prediction, *Bull. Am. Meteorol.. Soc*.,
280 doi:10.1175/2010BAMS2962.1.
281 Taylor, K.E., R.J. Stouffer, and G.A. Meehl, 2012: An overview of CMIP5 and the
282 Experiment Design. *Bull. Amer. Meteorol. Soc*., **92**, 485-498, doi:
283 http://dx.doi.org/10.1175/BAMS-D-11-00094.1.
284 Vimont, D. J.. 2005: The contribution of the interannual ENSO cycle to the spatial
285 pattern of decadal ENSO-like variability. *J. Climate*, **18**, 2080-2092.
286 Vimont, D. J., 2012: Analysis of the Atlantic Meridional Mode Using Linear Inverse
287 Modeling: Seasonality and Regional Influences. *J. Climate*, **25**, 1194-1212. doi:
288 10.1175/JCLI-D-11-00012.1
289 Zanna L., 2012: Forecast Skill and Predictability of Observed Atlantic Sea Surface
290 Temperatures. *J. Climate*, **25**, 5047-5056.

291

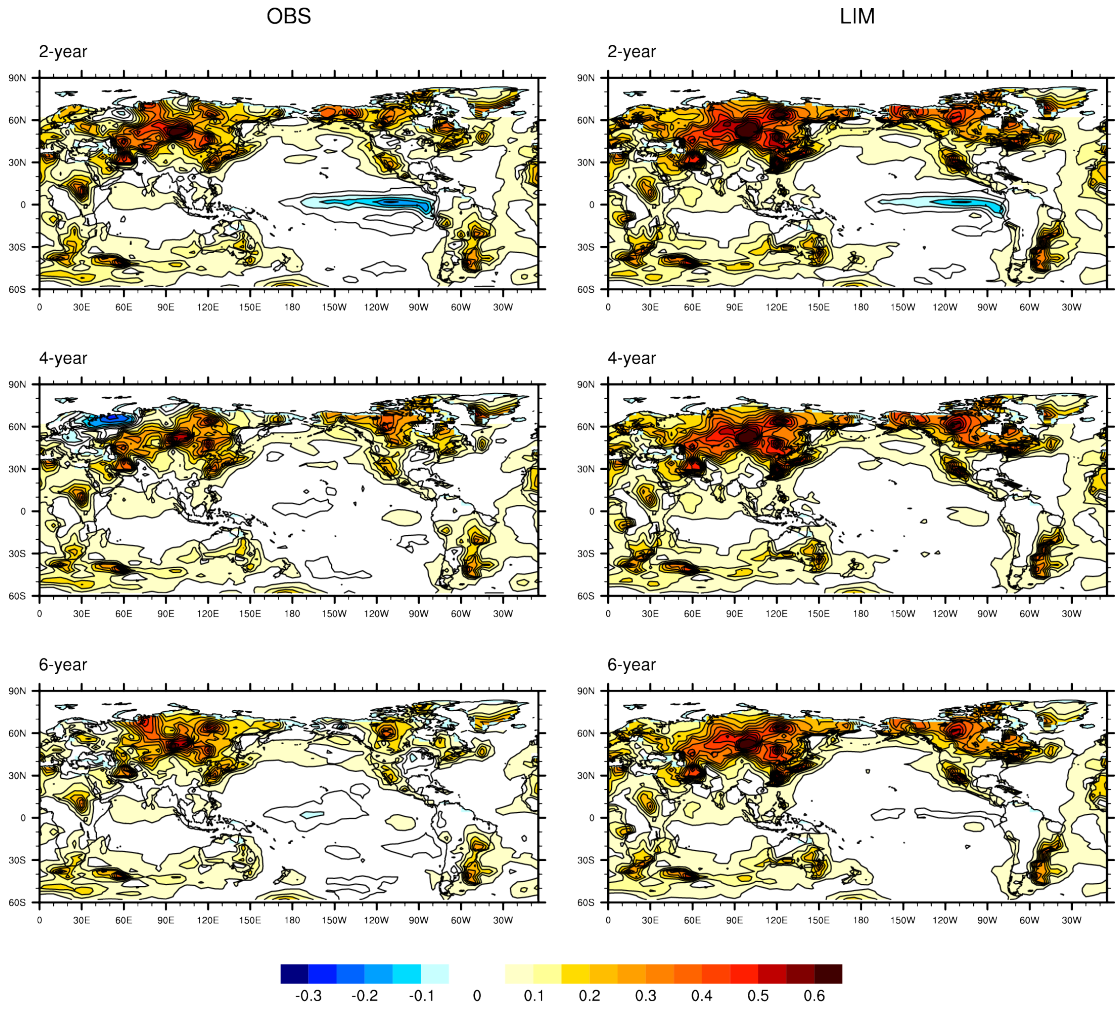


Fig. 1. Observed (left) and empirical model (right) surface temperature lag-covariance for lags of (top) 2 years (middle) 4 years and (bottom) six years. Contour interval is 0.05 K^2.

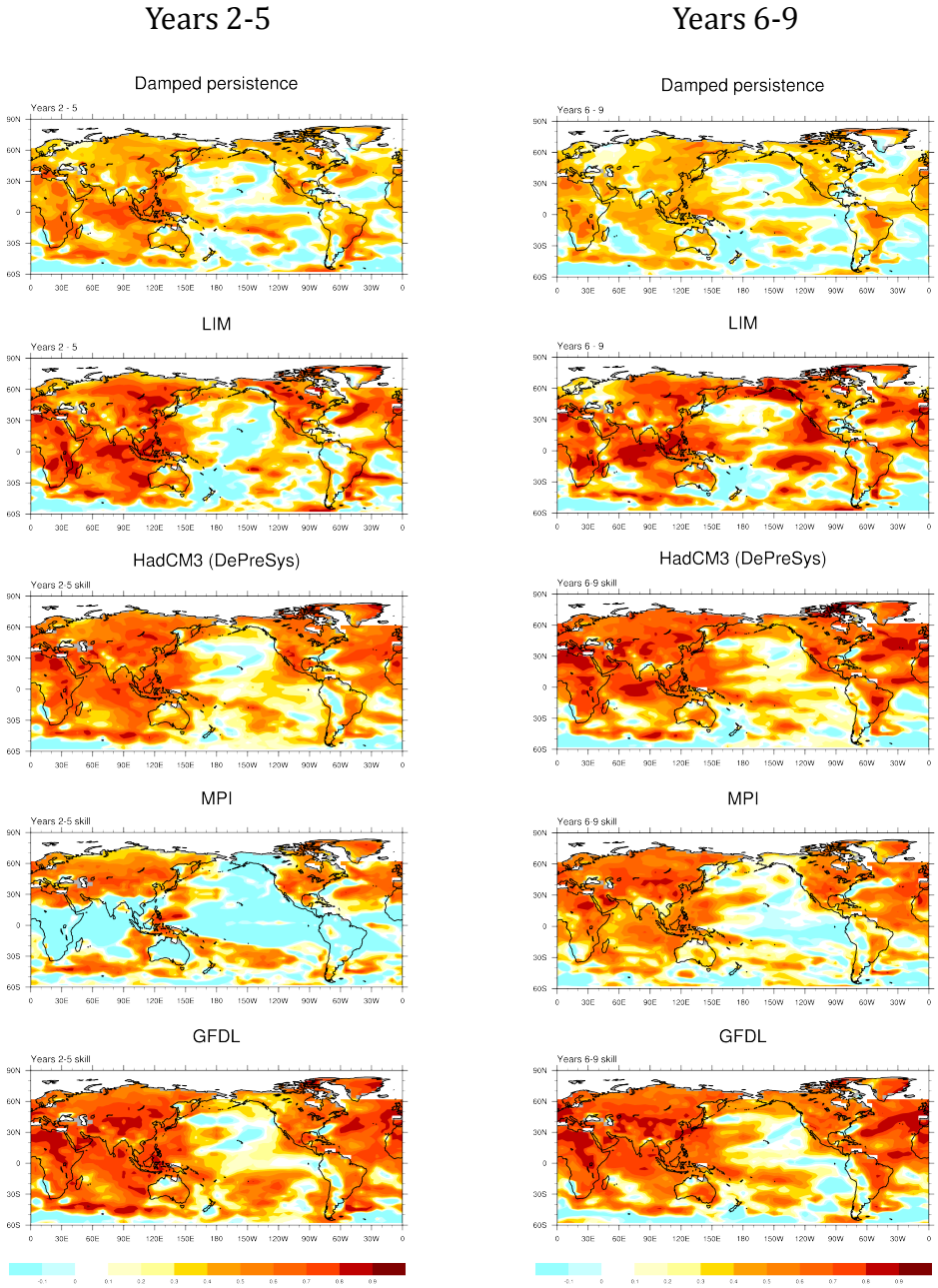
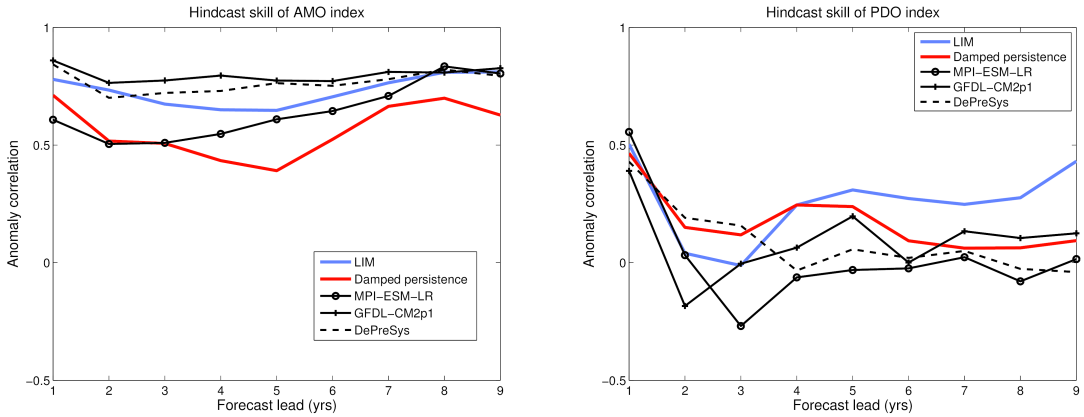


Fig. 2. Local anomaly correlation of (left) years 2-5 and (right) years 6-9 hindcasts for the CMIP5 models compared to damped persistence and the empirical multivariate AR1 model, for hindcasts initialized yearly from 1960-2000. (a) Damped persistence (b) empirical multivariate AR1 model (LIM) (c) HadCM3 (d) MPI-ESM-LR (e) GFDL-CM2p1. Contour interval is 0.1 with negative values indicated by blue shading. Shading of positive values starts at 0.1; redder shading denotes larger values of correlation.



304

305

306 Fig. 3. Skill comparison for the PDO and AMO indices from hindcasts initialized in the
307 years 1960-2000, calculated as described in the text. (left) AMO (right) PDO

308

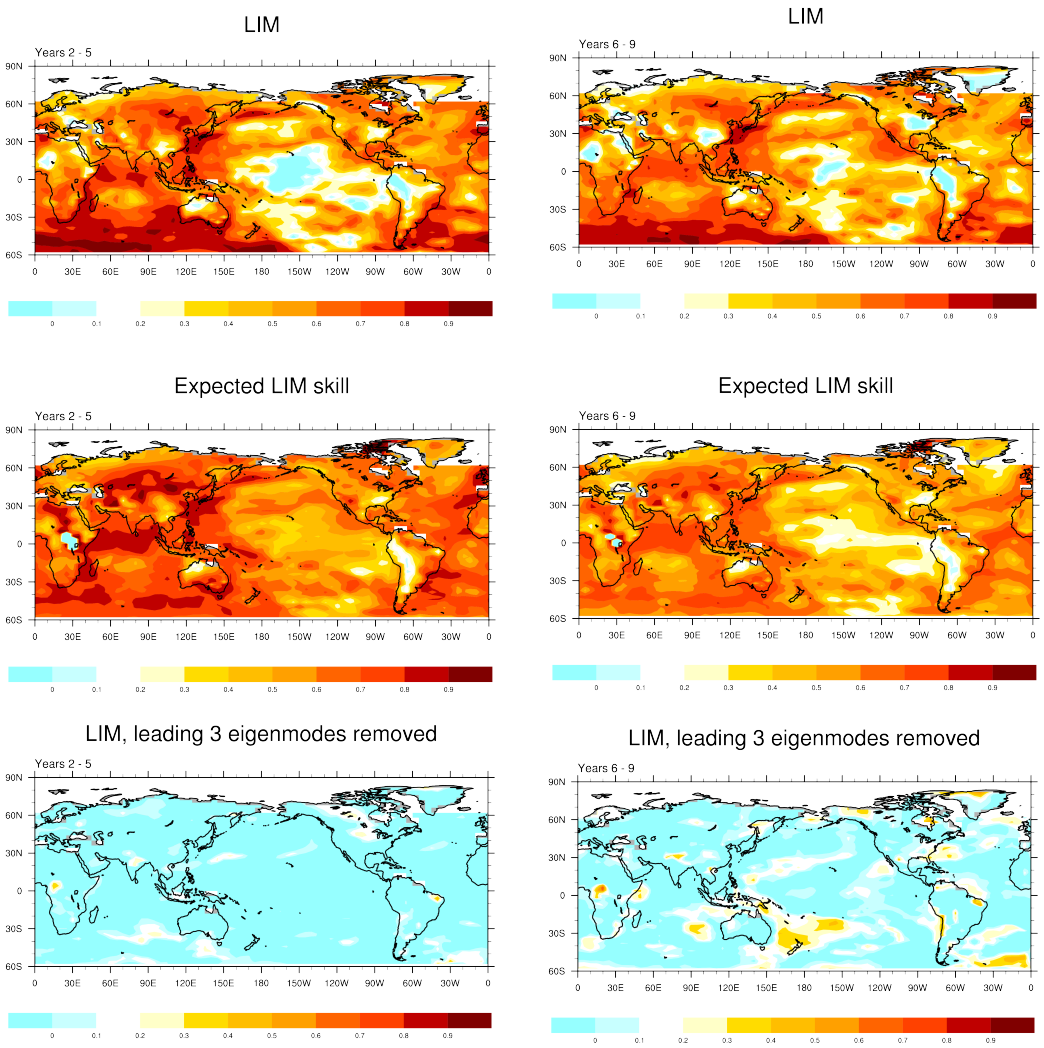


Fig. 4. Top: LIM skill for the 1900-2008 period for forecast leads of (left) 2-5 years and (right) 6-9 years. Middle: Same but expected skill. Bottom: LIM skill for the 1900-2008 period but where projection of the initial conditions on the leading eigenmodes (Fig. 5) are removed.

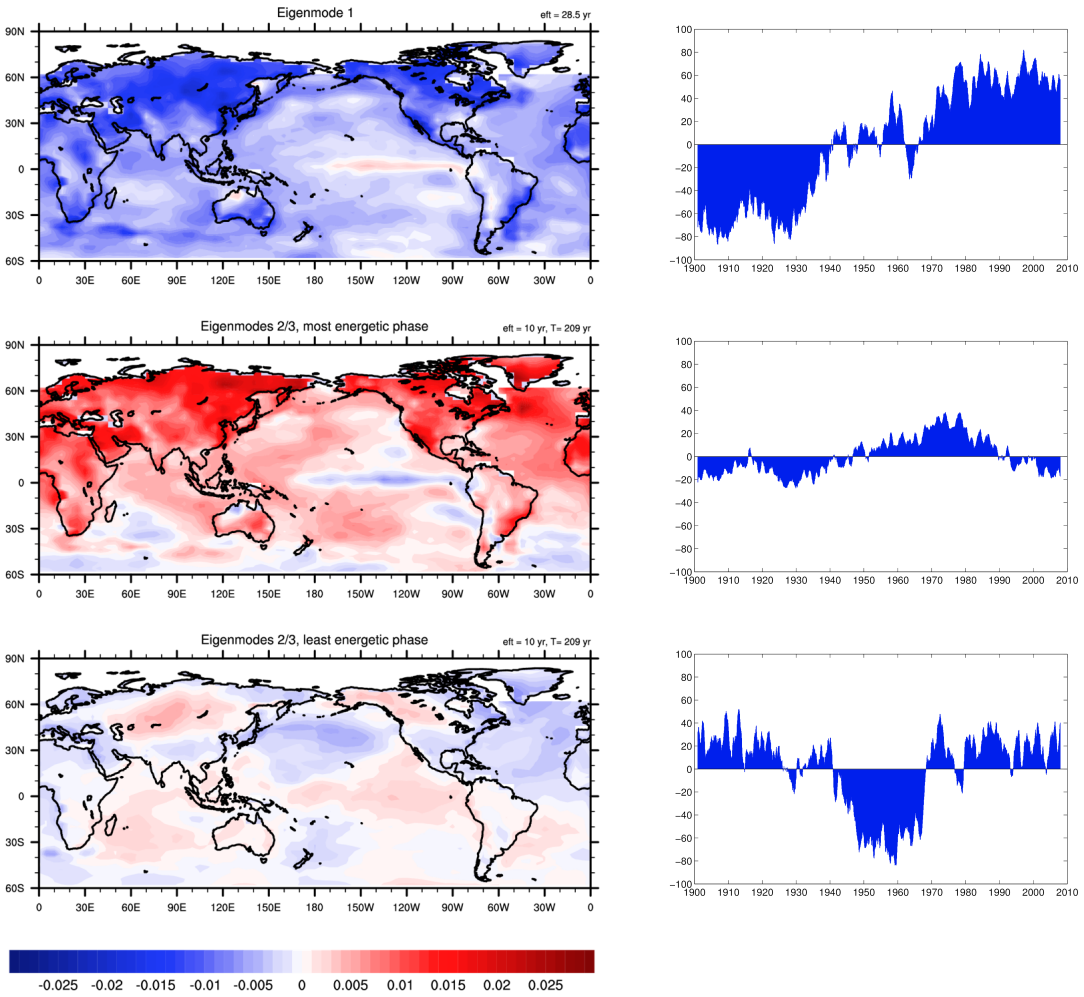
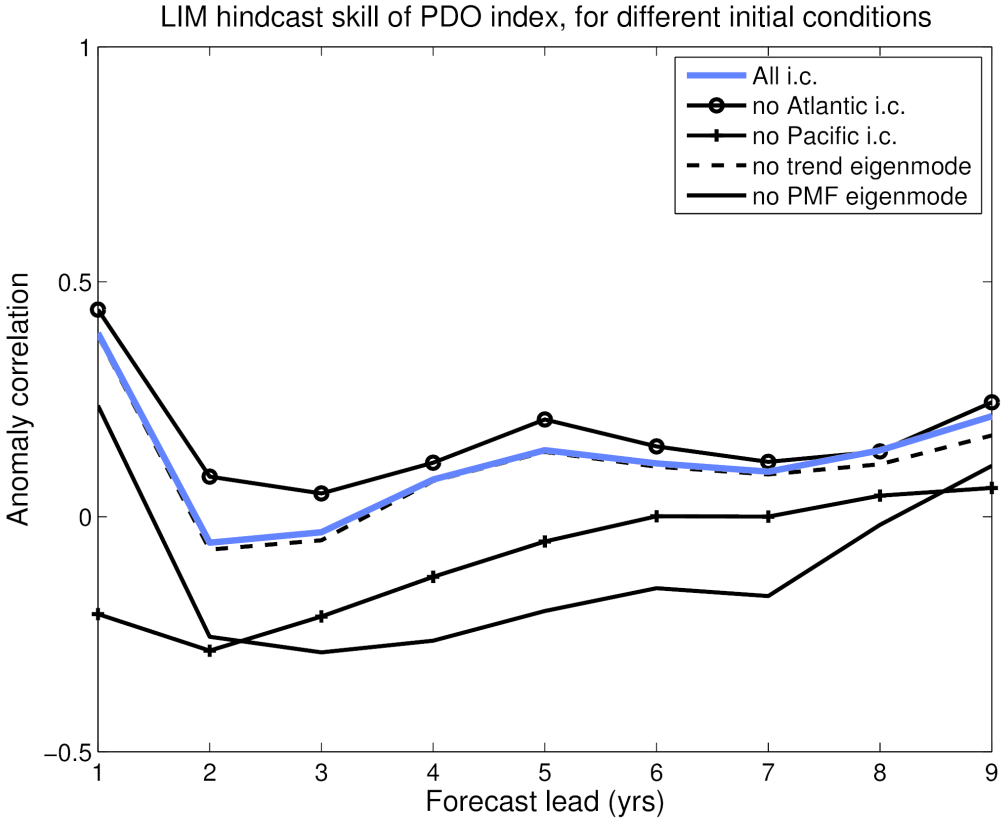


Fig. 5. Leading empirical normal modes, with their associated projection coefficient time series. Contour interval is the same in all panels. Sign is arbitrary but is consistent with coefficient time series. Red shading indicates one sign, and blue shading indicates the other sign.

326



327

328 Fig. 6. LIM skill of the PDO index, for hindcasts where different initial conditions are
329 used, for the 1900-2008 period. See text for description.

330

331

332